DOCUMENT RESUME

ED 461 263 FL 024 543

AUTHOR	Thain, John W.
TITLE	Improving the Measurement of Language Aptitude: The Potential Contribution of L1 Measures.
PUB DATE	1994-09-00
NOTE	39p.; In: Language Aptitude Invitational Symposium Program Proceedings (Arlington, VA, September 25-27, 1994); see FL 024 538.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Admission Criteria; Anxiety; English; Federal Programs; Grammar; *Language Acquisition; *Language Aptitude; Language Research; Language Tests; Learning Strategies; Listening Comprehension; Measurement Techniques; *Predictor Variables; Second Language Learning; Second Language Programs; *Second Languages; Stress Variables; Student Characteristics; Student Motivation; *Test Construction; Test Use; *Testing
IDENTIFIERS	*Defense Language Aptitude Battery; *Defense Language Institute CA; Noncognitive Attributes

ABSTRACT

Two studies at the Defense Language Institute (California) investigated the contribution of several variables to prediction of post-language-training proficiency: (1) scores on a general vocational aptitude battery and a language aptitude battery, both used to screen potential students; (2) scores on other cognitive measures not used in the screening process; and (3) scores and ratings on measures of student motivation, anxiety, and use of learning strategies. Two additional studies continued the effort to add certain types of native language competency measures to the Defense Language Aptitude Battery used for student selection. One competency measure considered was listening assessment, particularly as it accounted for two factors affecting the difficulty of listening tasks: the extent to which the examinee had the opportunity to rehearse the initial stimulus or recode it for later use; and the extent to which the examinee had a pre-existing mental set enabling application of an appropriate schema to select and organize the stimulus input as needed to perform the testing task. The second native language competency measure considered was grammar testing, particularly speeded grammar tests in which the task is to identify grammatical errors in sentences. Findings and implications for language aptitude test battery development are discussed. (Contains 113 references.) (MSE)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
John Thain
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

IMPROVING THE MEASUREMENT OF LANGUAGE APTITUDE: THE POTENTIAL CONTRIBUTION OF L1 MEASURES

John W. Thain
Defense Language Institute

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 461 263

Overview

Background

This paper begins with a brief sketch of work done in the area of language aptitude measurement at the Defense Language Foreign Language Center (DLIFLC) in the past eight years. There is no effort to go into detail into this sketch; however, the reader interested in further detail is provided with ample references to other presentations at this symposium and to other published works in the footnotes and bibliographic references following this paper. This cursory introduction does, however, define the instruments and measures used to screen potential students applying for language training at DLI. The intent is that this sketch will help provide context and points of orientation for the reader later on in this paper.

Adding L1 Measures as Predictors

The rest of this paper addresses the feasibility of adding two specific L1 measures as additional predictors to the current Defense Language Aptitude Battery (DLAB). DLAB is one of the batteries used to screen applicants for language training at DLI. The two potential predictors are (1) a test of L1 (native-language) listening comprehension and (2) a test of sensitivity to English grammar and usage.

Most of the paper deals with only one of the two potential predictors, an L1 measure of listening comprehension. The second potential predictor, a test of sensitivity to English grammar, is discussed only briefly.

The Main Body of the Paper: L1 Listening Comprehension as a Predictor

Five sections on listening comprehension in this paper

The part of the paper dealing with native-language listening comprehension can be further subdivided into five sections. The first section reviews the kinds of native listening comprehension (NL) tests currently available as models. The next three sections address several theoretical issues involved in the addition of NL tests to the current DLAB. The last section lists conclusions and recommendations.

Importance of the middle three sections

The content of the middle three sections mentioned above deserves further comment. There is little precedent for using NL tests as foreign language aptitude tests. A literature search was needed to address the relevant theoretical issues in using NL tests. I found three approaches in the literature that were relevant to the question of using NL tests as language aptitude predictors. Each approach was represented by its own literature, but no previous attempt had been made to synthesize information from these three perspectives to address the specific problem at hand. I call the three perspectives (1) the predictive perspective (2) the linguistic content perspective, and (3) the perspective of cognitive models. I needed not only to review three different kinds of literature, but in a sense, to attempt an unprecedented synthesis of three types of literature for a particular purpose. Hence, there needed to be three middle sections under the general topic of NL comprehension, each section concerned with one of the three approaches.

FL024543

BEST COPY AVAILABLE

The First of Three Approaches in the Literature on Listening: the Predictive Perspective

I call the first approach the predictive perspective. In the section dealing with this perspective, I refer to studies of the statistical characteristics of currently used screening measures, and the potential consequences for overall prediction of adding additional predictors. I address the effect of covariance between predictors on the total predictive power of a battery. I also mention the consequences of adding predictors that may themselves be multidimensional to existing predictors in a battery.

The Second of Three Approaches in the Literature on Listening: the Linguistic Content Perspective

I call the second approach the linguistic content perspective. The discussion of this perspective is more lengthy and complex than the discussion of the other two perspectives.

I point out that FL (foreign language) listening proficiency is one of the proficiency criteria we want to predict. I note how the concept of language proficiency in all skills, including listening, as expressed by the Interagency Language Roundtable (ILR) proficiency level scale, has been influenced by very basic, important, and overwhelmingly positive theoretical developments in the field of foreign language teaching methodology over the years. In the course of these developments, the ILR proficiency levels have established their unquestioned legitimacy as training criteria within the government and a large part of the progressive academic teaching community.[1]

Two consequences of the broad range of ability encompassed in ILR scales.

The ILR listening scale attempts to quantify a very broad range of proficiency. The lower part of the scale describes beginning language learners and the upper part of the scale describes polished bilinguals. This enormous range of individual differences seems to bring about two consequences.

Consequence number one. The first consequence is that different aptitude predictors may represent abilities that contribute in different magnitude at different levels of proficiency acquisition (and thus at different points on the ILR listening scale). I also note that the listening literature suggests there may be two types of listening, and that these two types of listening may make different cognitive demands on the listener. Each type of listening may have its own unique pattern of relationships with the other ILR skills. I conclude that evidence of multidimensionality in listening and of complex interrelationships among ILR skills could have interesting consequences for predictor-criterion relationships.

Consequence number two. A discussion of the first consequence leads us naturally to the second consequence. The ILR scale is "a "vertical" scale rising from Levels 0 to 5, a very great range of the ability. NL research looks at listening from a "horizontal" view that intersects only the top of the vertical ILR scale. Factors such as grammar, vocabulary, and phonology that play a major role for beginning FL listeners play a much lesser role in NL. In turn, NL research has identified separate listening factors that contribute to individual differences among native listeners, and these factors do not correspond to the factors contributing to individual differences at lower levels on the ILR scale.

Pure traits (PTs) vs. native authentic listening (NAL)

The difference between the "vertical" and "horizontal" perspectives is highlighted as I cite the work of the NL researchers Bostrom and Waldhart (1981). They resolved NL into three factors: (1) short-term listening (2) long-term listening (3) interpretive listening (sensitivity to affect).

---

[1] There are other scales for rating proficiency that are based on level systems similar to that used by the ILR. Examples include the ACTFL scale used by the American Council of Teachers of Foreign Languages and other rating scales used in Europe.

I contrast a view of NL based on a three-factor analysis similar to that of Bostrom and Waldhart, and a "global" view of NL as "native authentic listening" (NAL). After I coin a term by calling each factor in the three-factor analysis a "pure trait" (PT), I broach an important question that I do not immediately answer: "Is a NL test based on PTs a better predictor of ILR proficiency levels than a test based on NAL?"

The perspective of cognitive models

I call the third perspective the cognitive modeling approach. I sketch evolutionary changes in the field of psychology from radical "black-box behaviorism" days to current day cognitive psychology, including the development of the field of artificial intelligence (AI). AI specialists have successfully modeled human comprehension of language. Within a limited range of topics, machines can now carry on reasonable conversations with humans in which they make many of the inferences that humans would make in similar circumstances.

In the context of these developments in AI, I draw a series of analogies between listening comprehension and the operation of a multimedia database. I point out that the series of analogies leads to conclusions similar to those of Bostrom and Waldhart concerning the multidimensional nature of NL.

Conclusions and recommendations for further study about listening comprehension

The last section on listening comprehension lists conclusions and recommendations. I list a set of criteria for evaluating possible listening comprehension measures for inclusion into the DLAB. I categorize the NL tests reviewed earlier in terms of whether they measure PTs (pure traits), native authentic listening (NAL), or some mixture of the two. I then list some of the issues in using PTs as language aptitude measures, and related issues in using measures of NAL as language aptitude measures.

The Rest of the Paper: Tests of Grammatical Sensitivity as Predictors

In the last part of this paper, I review tests of grammatical sensitivity, but not in the same detail with which I reviewed NL tests earlier. Two types of tests are reviewed: (1) tests of sensitivity to English grammar, and (2) tests of sensitivity to foreign (or artificial) language grammar rules.

Overview of organization of the paper

- This overview spans pages 1-3 of the paper.
- A sketch of background information and references to related presentations at this symposium are to be found at pages 4-6.
- A major division of the paper entitled "Exploring Native Listening Comprehension" spans pages 6-28.
    - A review of currently available NL comprehension tests is found at pages 6-7 under the main division heading.
    - A section entitled "First of Three Complementary Approaches: the Predictive Perspective " spans pages 8-9.
    - A section entitled "Second of Three Complementary Approaches: the Linguistic Content Perspective "spans pages 9-16.
    - A section entitled "Last of Three Complementary Approaches: the Predictive Perspective:" spans pages 17-23.
    - A section entitled "Conclusions and Recommendations Concerning NL Measures" spans pages 23-28.
- A major division of the paper entitled "Exploring Tests of Grammatical Sensitivity in English" covers pages 29-32.
- Bibliographic references are found at pages 32-38.

Background Information and Related Presentations at this Symposium

General

A major study conducted at the Defense Language Institute (DLI) from 1986 to 1989 investigated how well a variety of variables predicted proficiency after language training. This study, the Language Skill Change Project (LSCP), was a longitudinal study designed to follow approximately 2,000 Army "linguists" throughout a four-year period. Data collection points included (1) initial aptitude screening prior to entry into the Army; (2) several occasions in the course of language training,; and (3) post-graduation field assignments. The population sample included both students of Spanish, German, Russian, and Korean. Another presentation describes this population sample in more detail.

A secondary study used a portion of the same data base to investigate predictors of attrition from DLI training.

In both the longitudinal and the attrition studies, the predictor variables used to predict language training success included (1) scores on a general vocational aptitude battery and a language aptitude battery, both used to screen potential students; (2) scores on other cognitive measures not used in the screening process; and (3) scores and ratings on measures of student motivation, anxiety, and use of learning strategies.

Criterion measures included (1) successful course completion as opposed to attrition from training; and (2) the Defense Language Proficiency Tests in these languages for speaking, listening, and reading skills.

Aptitude Variables used in Official Screening

Aptitude tests used in official screening are not administered by the DLI. These tests are normally administered by the interservice Military Enlistment and Processing Command (MEPCOM).

Applicants for military service must attain passing scores on a composite of the Armed Services Vocational Aptitude Battery (ASVAB), a paper and pencil general vocational aptitude battery. The passing scores have hardly changed since the LSCP was conducted. ASVAB includes tests of verbal, mathematical, technical (mechanical and electrical), and clerical coding abilities. The verbal tests include measures of paragraph comprehension and vocabulary knowledge. All ASVAB test materials are printed in English.

Examinees reaching certain minimum scores in specified components of the ASVAB are eligible to take the Defense Language Aptitude Battery (DLAB). This battery contains several subtests.[2] The subtests measure (1) identification of syllable stress (2) deductive language learning of an artificial language (3) inductive language learning from pictures and artificial language work sample. The first two subtests are presented on tape, and the third subtest is printed in the test booklet.

Scores on the ASVAB and DLAB tests administered by MEPCOM would normally be present in the official personnel records of students even before students arrive at DLI.

Other Cognitive Measures Not Used in Official Screening

After completing basic training, the students in the LSCP sample actually arrived at DLI. DLI administered additional cognitive tests to them as part of the LSCP. These tests included the Watson-Glaser Critical Thinking Appraisal, the Flanagan Expression Test, and the Flanagan Memory Test.

---

[2]See reference by Petersen, C., Al-Haik, A. (1976)

Measures of Student Motivation, Anxiety, and Learning Strategies

In order to assess motivation to learn a foreign language immediately prior to language training, the subjects were administered Gardner Questionnaire Form A. This questionnaire was a modification of previous questionnaires used by Gardner in earlier research and included scales for Integrativeness, Instrumental Motivation, and Interest in Foreign Language. [3]

During the course of language training, Gardner Questionnaire Form B was administered. This questionnaire included scales for Motivational Intensity, Attitude Toward Learning, Class Anxiety, Use Anxiety, Desire to Learn, Attitude Toward the Instructor, Attitude Toward the Course.

The Strategy Inventory for Language Learning (SILL) was also administered to measure self-reported use of learning strategies during instruction.

Results of Background Studies

The results of these studies have been already described in another paper at this symposium.[4]

In the basic study, stepwise multiple regressions with forced order of entry indicated that (1) general vocational aptitude (measured by ASVAB), (2) language-learning aptitude as (measured by DLAB), (3) measures of student motivation, anxiety, and learning strategies use, (4) additional cognitive measures not included in the official screening process all added contributions to predictive power. However, the pattern of multivariate prediction varied across the four languages taught and across the three criterion language skills .

A secondary study used a restricted set of variables. Course completion (as opposed to attrition from training) was used as a criterion measure. Chi-square interaction analyses (CHAID) indicated that (1) the pattern of interaction of variables varied across languages (2) both DLAB and the additional cognitive measures not included in the screening process contributed to the segmentation of subsamples. The subsamples in individual languages were segmented on the basis of the differentiating criterion of percentage of successful course completion.

Related studies and follow-up studies

Shortly after the above mentioned studies were completed, DLI launched several simultaneous efforts to improve aptitude prediction: (1) an item analysis of the current DLAB (2) an effort to compare languages in terms of the "factors" that made some languages more difficult to learn than others (3) an effort to specify the kinds of language abilities and measures that should be included in an aptitude battery.

The results of the item analysis of DLAB were reported in another presentation at this symposium.[5]

Another presentation at this symposium addressed the second and third efforts mentioned above.[6]

---

[3] See reference by Gardner, R., Lalonde, R., Moorcraft, R., Evers, F. (1985).

[4] "The Defense Language Aptitude Battery: What is it and how well does it work?", by John Lett and John Thain.

[5] "The Defense Language Aptitude Battery: What is it and how well does it work?", by John Lett and John Thain.

[6] "Psycholinguistic Issues in the Assessment of the Subcomponents of Language Abilities, by Brian MacWhinney.

Conclusions drawn concerning possible addition of L1 measures to DLAB

As noted above, the current DLAB contains test items based on artificial language material. This material taps primarily grammar learning and grammar analysis abilities. It does not contain test material based on normal L1 (English) language.

The other battery used in official screening process; the ASVAB, does include written L1 (English) tests of verbal ability, but does not include auditory tests.

DLI staff examined all of the information from the LSCP data base and recommendations resulting from the follow-on work mentioned above. DLI then decided to explore the possibility of adding two additional predictors to the language aptitude battery: (1) an L1 native speaker (English) test of listening comprehension (2) a test of sensitivity to English grammar and usage.

Exploring Native Language Listening Comprehension

Review of native-language (NL) listening tests

Introductory comments

I began my exploration of L1 native listening comprehension as a potential predictor by reviewing English native listening (NL) tests.

I discovered that NL test developers tended to see NLs as listening "skill-users" with a function and corresponding work to do in the native society. These developers perceived the NL as a student, teacher, counselor, or businessman; they felt his function was to learn, to help others, or to serve as an employer. NL test developers differ from FL test developers in this respect. They show less interest in clearly separating "language listening skills" from other useful skills and knowledge.

I quickly detected something interesting about English listening comprehension testing of *foreign students at English-speaking universities*--namely the tests used had more in common with NL tests of listening than with tests of foreign language (FL) listening comprehension. For this reason, we included such listening tests in our review.

I also found another interesting difference between contemporary NL and FL listening testing and research. Nowadays many FL testers, especially those at federal government institutions, want to test "proficiency," i.e. authentic and useful language. They don't want to test anything that looks like a classroom drill or an isolated piece of language. On the other hand, NL researchers are showing interest in testing memory span for letters and similar tests of short term memory.

While NL testers may concede such skills may be not useful in isolation, they tend to find these measures to be useful as (1) predictors of more complex behavior, or (2) moderating variables for cognitive models of more complex skills, or (3) diagnostic devices.

NL Tests Reviewed

I reviewed seven tests which I found to be mentioned in the literature. Brief synopses follow:

Watson-Barker Listening Comprehension Test. This test includes subtests for "listening to a lecture," "emotional listening," "instructions and directions," "listening for content," and "listening to conversations." Businesses have used this test to accompany training programs. The University of Illinois has used it to differentiate levels of listening skills of foreign students taking classes at the University. The test is presented by means of videotape. The publisher is Spectra Communications in New Orleans, LA.

Kentucky Comprehensive Listening Test. This test is a multiple choice test with four parts. The four parts measure performance on the following tasks:

(1) Listening to letters or number strings amidst distracting noise. The examinee is prompted immediately after the stimulus to identify the relative position of a letter or number in the string.

(2) Listening to letters and number strings without the presence of distracting noise, but with a delayed prompt to identify the relative position of a letter or number.

(3) Listening for real meanings (i.e. illocutionary acts) hidden in very short answers in a dialogue with strong nonverbal affective signals

(4) Listening to a 1500 word lecture.

The publisher is the Kentucky Listening Research Center in Lexington, Kentucky. Data collected on this test are particularly interesting (1) because it has been used in a variety of research and practical contexts, (2) the authors have fostered a series of studies from which a particularly fruitful nexus of explanatory constructs has evolved.

Carleton University Test.[7] Carleton University in Ottawa, Ontario, has constructed a listening test that it administers to its incoming foreign students. The examinees take notes on a lecture, actually reorganize their notes, and then do library research on the basis of their reorganized notes. The criterion for success is the quality of their library research. Test results reportedly correlate highly with an English comprehension test developed by the University of Michigan.

NTE Core Battery Test of Communication Skills. The National Teacher's Examination (NTE) program includes a Test of Communication Skills, which includes subtests in listening, reading, and writing. Many sample listening items given in the test information brochure are based on typical listening comprehension situations. However, item content is biased toward typical situations in which teachers might be involved. Some of questions defining the examinee's task include: (1)"Why does the man hesitate to call William's parents?;" and (2) "What assumption does the speaker make about high schools?"

The NTE School Guidance and Counseling Examination. This test includes a listening component, which is administered as part of a larger battery. The battery as a whole evaluates the skills and knowledge required of school counselors. In this test, the examinee listens to test items depicting situations in which counselors may be involved. The examinee then answers multiple-choice items introduced by item stems such as "The client is likely to react by..." or "The counselor's objective was..."

Brown Carlsen and STEP. Two older NL tests include (1) the Brown-Carlsen Listening Comprehension Test, from Harcourt Brace and Jovanovich, and (2) the STEP (Sequential Tests of Educational Progress) Listening Comprehension Test, once published by a since dissolved ETS subsidiary. The Brown-Carlsen test has subscales that measure vocabulary, recognition of transitions, ability to follow directions, immediate recall, and retention of facts from a lecture. The STEP listening test was one of seven tests in a battery, which included tests of reading, spelling, and other achievement areas. It was published in a series of forms that spanned grade levels 4-14.

Introducing Three Complementary Approaches in the Literature on Listening Comprehension

In a general sense, there is an abundance of literature on NL. On the specific point of view of use NL as a predictor of foreign language proficiency, there is a poverty of literature.

The general literature on NL suggested several complementary perspectives for understanding the subject area. I became aware that many people in the field of language aptitude measurement may seldom have considered these perspectives about NL *in conjunction with each other*. I believe the approaches are

---

[7] Personal communication from Janna Fox at Carleton University. See also reference by Janssen, C., Hansen, C., Buck, G., DesBrisay M., Fox, J., Shohamy, E., (1993).

synergetic. This means that insights and conclusions gained from one perspective can influence one's thinking in following up other approaches. One of my objectives is to improve communication between investigators using different approaches and to stimulate discussion about new ideas arising from the interaction of approaches.[8] I will first touch on several seemingly loosely related ideas, and then attempt to tie them together with some concrete examples.

I have called three of these diverse points of view the (1) predictive perspective; (2) the linguistic content perspective; and (3) the cognitive model perspective.

First of Three Complementary Approaches: the Predictive Perspective:

The general standard regression formula for prediction is:

$$Y = \sum_{i=1}^{n} \alpha_i x_i + C; \quad \text{where } \alpha_i x_i \neq 0, \quad n \geq 1.$$

In this general formula, Y is the criterion, and i is the number of predictors contributing to the equation. Each of the n predictor values is multiplied by its own weight α_i and then all the weighted predictors are summed to give the overall weighted contribution of all the predictors in the equation. The values of the weights are affected by covariance between the predictors. This general formula can apply to the prediction of any proficiency criterion from any number of NL predictors.

The mathematics of prediction are straightforward. However, communication problems can arise among investigators with different backgrounds for reasons that have little to do with the mathematics of prediction. For this reason, in the following paragraphs I will be trying to accomplish two things at once. I will list the possible predictors that might go into a predictive equation, but at the same time I will also be explaining how researchers with different perspectives might have divergent views on how many predictors should be in the equation, and how these predictors are interrelated.

L1 Predictors already included in general aptitude batteries even before language aptitude testing.

In the case of the Defense Language Institute, a passing score on a general aptitude battery, the ASVAB, is a prerequisite for taking the DLAB. Hence, there are already some potential predictors from ASVAB available for inclusion in the equation above (before considering any specific FL aptitude predictors or any new potential L1 predictors.) General aptitude tests such as the ASVAB typically include subtests that represent the V (Verbal) factor as well as other familiar factors such as the N (numerical) factor.[9]

---

[8] Some of these approaches may seem on the surface to diverge from the ideas underlying our use of the ILR proficiency scale as criteria. Where this may seem to be the case, I will pause to explain exactly what elements of these approaches I find useful and compatible with the ILR approach.

[9]Other factors in ASVAB (or similar general aptitude measures) besides the V factor are likely to contribute to the prediction of language proficiency. The V factor is mainly relevant to the discussion here in this section, because this section focuses on L1 comprehension measures. For more detail, see references by Silva, J., White. L. (1992); Department of Defense (1985); Kass, R., Mitchell K., Grafton, F., Wing, H. (1983); Carroll, (1958); Carroll (1962), Carrol (1993). Tests that consistently correlate with each other more than with other types of tests are assigned to the same "factor". The "V" factor is consistently represented by L1 vocabulary tests. There is no hard and fast theoretical reason in the field of psychology that a "V" factor should be exclusively identified with any of the four L1 skills. However achievement and aptitude batteries, including ASVAB, normally include a reading comprehension test; but they include tests of the other three skills less often. It is easier to produce, administer, and score multiple-choice reading

Additional predictors in language aptitude batteries not identifiable with any of the four skills

A variety of studies have identified factors related to phonology, grammatical sensitivity, and word association that contribute independent variance toward the prediction of L2 Proficiency beyond that contributed by the "V" L1 factor included in general aptitude batteries. However,-- (1) the "V" factors, (2) these additional aptitude factors, and (3) any additional L1 predictors we may choose to add--may *all* share some covariance. This covariance would (1) affect the weights in the prediction equation so that all weights would have to be recomputed with the addition of each new predictor, and (2) tend to limit increases in the size of a multiple correlation coefficient with the addition of each predictor, (to the extent that each predictor added shared variance with predictors introduced earlier in the equation.)

Research on Potential L1 Listening Predictors that lack parallelism to ILR/ACTFL criterion scales

FL researchers using the ILR and ACTFL proficiency scales as criteria tend to consider L2 listening as a unitary trait. They may tend to assume that NL listening would also be an unitary trait. If NL listening were a unitary trait, a single additional predictor would be added to the equation to join the predictors mentioned earlier.

However, a contrasting perspective will be discussed later in this paper. At that time, I will point out that two prominent NL researchers have attempted to analyze NL into three component traits. Users of the ILR/ACTFL scale may be forewarned that only one of these traits bears some similarity to the kind of global "listening" with which they are familiar. My interest is in these NL component traits as potential NL *predictors* of ILR proficiency levels, *not* as *alternatives* to the ILR listening proficiency *criteria*.

Concepts of FL listening that are different in emphasis from the ILR/ACTFL perspective

In addition to the NL researchers discussed in the previous paragraph, there are some writers on FL listening who at times don't see the four language skills as distinct "points," as much as moist blurry ink blots that overlap each other. My interest is in how their insights shed light on the aptitude-proficiency (*predictor-criterion*) correlations across languages and language skills. The purpose is not to advance these ideas as alternatives to the ILR skill level *criteria*.

In the next section on the "linguistic content perspective," I will attempt to explore predictor-criterion relationships from a different point of view and attempt to bridge a communication gap between researchers with different points of view.

Second of Three Complementary Approaches in the Literature on Listening Comprehension: the Linguistic Content Perspective

Introduction.

There is another reason why foreign language listening and NL researchers might not initially communicate. NL researchers may not be very familiar with the development of FL teaching methodology in the past 75 years. For this reason, I will very briefly sketch how the relevant developments in teaching methodology may have shaped the ILR listening scale and ILR testing procedures. The intent is to establish the critical importance of the ILR scale as a foreign language criterion measure.

I then discuss alternative conceptualizations of the interrelationships among FL listening and other FL skills advanced by scholars not closely associated with the ILR testing community. The intent is to add

---

tests than tests of the other three skills. Hence psychologists tend to identify L1 reading comprehension almost as closely with the "V" factor as vocabulary tests.

a relevant perspective to our in-house thinking about skill prediction. Finally I contrast a generally accepted concept of FL listening with a "nonparallel" concept of "listening" advanced by two NL researchers.

From Grammar Translation to Proficiency

In 1930, the grammar-translation method was used to teach foreign languages in America. By 1960, the audiolingual method had displaced this earlier method. Since then, specialists in FL teaching methodology have been distancing themselves from both approaches.

They have begun to define authentic language use is the *ultimate* instructional goal. That is, the ultimate goal (perhaps unattainable by most second language learners) was that second language learners should read, speak, listen, and write languages the way native speakers use their language. In addition, the FL methodologists recognized and defined a variety of *intermediate* levels of ability for using and understand authentic language short of native speaker capability.

The FL methodologists rejected the grammar translation approach because they noticed that second language learners could learn grammar rules and do translations without much progress toward being able to read, speak, listen, and write languages the way native speakers used the language (or toward any recognized intermediate level of authentic language use).

They also rejected the audiolingual approach because they noticed that although second language learners could acquire habits for listening to and repeating small segments of language, they were not necessarily making progress in the sense of using progressively more complex cognitive processes through the new language.

The Proficiency Movement

Prominent FL methodologists joined together in the "proficiency movement." This movement included representatives from the structured, intensive language programs of the Federal Government and from a variety of language programs within academic institutions. Members of this movement began to define proficiency as their criterion goal. They defined a proficiency scale for each skill in terms of increasing ability to accurately use the new language to accomplish increasingly difficult authentic language tasks. All foreign language learners as well as native speakers were rated on a continuous scale across an enormous range of ability, from rank beginners to polished bilinguals--students and teachers alike.[10] Testing tasks and items showed a corresponding range of difficulty. At any given point on this broad continuum of item difficulty, it was assumed that an item on a listening proficiency test should be set in a meaningful situation in which language students might actually find themselves using the language in real life.

Today a progressive ILR tester in the proficiency movement may tend to consider it a throwback to obsolete unproductive methodology to include items in a FL listening test which consist of isolated bits of language (for example, isolated sounds or letters). Furthermore, such a tester might argue that a test of

---

[10] It is very important that there be a cooperative program between the government and academia to use compatible testing systems that measure such a broad range of ability. A teacher needs to master the language he teaches, and also master the language in which his/her employing institution imparts training in FL methodology (usually English in the United States). For these reasons, it is difficult to conceive an effective national level policy for fostering language training in this country without such a testing system. A testing system is needed to manage the career cycle of the two main classes of people who become foreign language teachers in America: (1) American-born language students who learn enough of a foreign language to be able to themselves teach foreign languages to other Americans; and (2) foreign-born teachers who first become students to learn English and then subsequently teach foreign languages to Americans. For more detail see references by Carroll, J. (1967); Higgs, T., Clifford R. (1982); Heileman, L., Kaplan, I., (1985), James, C. (ed.), (1985), Lowe, P. (1985), Clark, J. (1986), Child, J. (1987), Valdman, (ed.), (1987), Clark, J., Clifford R. (1987), Child, J., Clifford, R., Lowe, P. (1993), Hadley, A. (1993).

listening should not permit the examinee to answer a question or solve a communication problem without being forced to understand the lexis and grammar of a foreign language text (e.g. as would be the case if the examinee answered a question about text solely by correctly interpreting a combination of gestures and voice modulation or by relying on background knowledge and context.)

Effect of Unrestricted Range in the Population Tested on Observed Variance

When we FL researchers define a "listening" trait using a rating scale like the ILR scale for a broad population ranging from beginning learners to polished bilinguals, we find statistical evidence for considering "listening" a unitary trait. An overwhelming amount of variance is contributed by huge individual differences in mastery of foreign language codes. All other possible contributing traits are but drops in this vast ocean of variance.

The analogy of a vast ocean and vast variance can be extended further. ILR proficiency scales depend on individual differences in factors such as vocabulary, grammar, and sociolinguistic competence to discriminate among a great range of ability in the population.[11] The situation may be different for NL testing. In contrast, many differences in native listener performance may be less dependent on individual differences in vocabulary, grammar of the native language or knowledge of one's own native culture than on other factors. This suggests a way to complete the ocean analogy. If somehow all the water in the ocean evaporated, a theory based on the ocean being comprised of 96% water would not be a very good schema for making an inventory of the salts, minerals, fish, plants, and rocks left behind.

Good for the goose, but not for the gander

I hesitate to consider my experiences as an ILR proficiency tester as a warrant to evaluate the kind of issues that should be considered important in the field of NL testing (or specifically in the field of NL testing as a predictor for FL proficiency.) In this area, I believe ILR testing experience needs to be supplemented by perspectives from NL testing, and by other perspectives from the FL research community.

However, before proceeding to introduce some other helpful and complementary perspectives, let me hasten to preclude any misunderstanding based upon my previous statements. In general and for all practical purposes, I consider (1) that FL teaching methods have evolved in the right direction; (2) the concomitant trend toward accountability both in the government and in universities is good; and (3) our ILR criterion of "foreign language proficiency," specifically including listening proficiency, is defined properly.

An overview of other perspectives from the FL research community

Should skills be viewed as "distinct points" or "blurry inkblots?" Table 1 lists distinctions found in the literature that potentially cut across skills. The information in the table highlights the possibility that some types of L1 listening may make cognitive demands that are similar to those required in L1 speaking, while other types of L1 listening may make cognitive demands that are more like L1 reading.

If we plan to use L1 listening to predict L2, these distinctions are potentially important because the distinctions in L1 may have parallels in L2. Tannen's (1982) oral-literate style distinction may illustrate this point

---

[11] For a more complete elaboration of this point, see reference by de Jong, J. (1994). As de Jong points out, if one looks *closely* at *any* narrow subinterval on the broad scale of language proficiency (not just at the top of the scale for native proficiency as I am doing in this paragraph), one can probably find evidence for trait multidimensionality within that specific *subinterval*. On the other hand, if one takes a *broad* overview of the *whole* language proficiency scale (from a *distance* to use de Jong's metaphor), the scale as a whole appears to be unidimensional.

TABLE 1[12]
TWO TYPES OF LISTENING?
A CLASSIC EXAMPLE OF FUZZY SETS

SOURCE	MORE LIKE SPEAKING	MORE LIKE READING
Tannen (1982)	Oral style	Literate style
ILR	Street	School
ILR	Participatory	Nonparticipatory
Bostrom (1981)	Interpretive listening	Lecture listening
Cummins (1982)	Contextualized Requires BICS (Basic Interpersonal Communication Skills)	Decontextualized Requires CALP (Cognitive Academic Language Proficiency)
Canale (1982)	Interactive	Autonomous
Rost (1990)	Collaborative	Transactional
MBTI Thinking/Feeling[13]	Feeling type favored	Thinking type favored
Brain-hemisphere studies	Right brain favored	Left brain favored
Other	Situation-based	Idea-based
	Listener plans to politely clarify speaker's role, intentions, or feelings as part of listening process.	Listener plans to make mental or written notes as part of listening process, with the intention of later consulting dictionaries, textbooks, or other reference works.

Some measures of L1 listening may (1) be more closely related to L1 reading; (2) tend to covary with ASVAB, because ASVAB as a whole is probably more "literate" than "oral;" (3) tend to predict L2 listening skills that are more "literate" than "oral".

Other measures of L1 listening may (1) be more closely related to L1 speaking (2) tend to add distinct variance not already represented in ASVAB (3) tend to predict L2 listening skills that are more "oral" than "literate."

The above observations seem to have potential predictive consequences: (1) adding L1 listening predictors may improve prediction of other ILR skills than listening as much or more than these predictors

---

[12] It should be emphasized that the two types of listening implied by Table 1 above are classic examples of "fuzzy sets." The various distinctions listed cut across each other and overlap. For example, (1) some lecturers may use "oral" styles to better communicate technical information to their audience (2) some face-to-face speakers may address very technical or even esoteric subjects. (3) certain lecture and staff meeting settings may be viewed as continuous discourses in which the listener shifts back and forth from a nonparticipatory status to a participatory status (as in question and answer sessions after lectures, or in briefings from individual departments in the course of some staff meetings) (4) certain interactive situations could place demands on the "thinking," "left brain", "idea-oriented" side of the listener, while certain noninteractive situations could place demands on the "feeling", "right brain", and "people-side" of the listener. Although the list of fuzzy points admittedly could be extended indefinitely, I still think there seems to be enough of a pattern present to talk about two "fuzzy sets" rather than a list of totally random and unrelated distinctions.

[13] See reference by Myers, J., McCauley, M. (1985).

may improve prediction of listening itself (2) it may not be possible (or even desirable) to have a neat paradigm of L1 skill predictors that match corresponding L2 skill criteria.

A perspective for viewing predictor-criteria interactions. Figure 1 portrays the kind of predictor-criterion relationships suggested in the previous section. The right and center portions of the diagram essentially carry over information introduced in Table I. The left side of the diagram contains new information. It depicts the three skills Speaking (S), Listening (L), and Reading (R) as irregularly shaped forms in definite spatial relationship to each other.

S is portrayed in a shape like a catcher's mitt, L in the shape of a peanut, and R is shaped like a feather. The upper part of the catcher's mitt S encloses the upper part of the peanut L. The lower part of the peanut L impinges on the upper part of the feather R. The lower part of the catcher's mitt S curves toward the lower part of the peanut L and the base of the feather R.

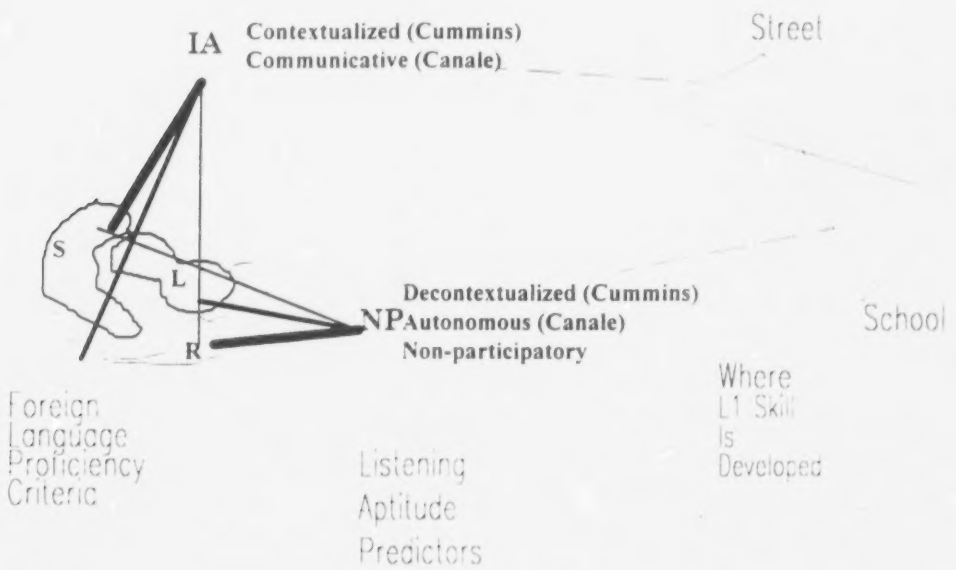


Figure 1

L1 LISTENING AS A PREDICTOR OF L2 PROFICIENCY SKILLS

The following analogies can be drawn. The upper parts of S and L approach each other; this symbolizes the close interaction between S and L in interactive settings. The lower part of L and the upper part of R approach each other; this symbolizes the textual similarities found when listening to formal lectures and reading subject matter texts.

The lower part of the catcher's mitt S approaches the lower part of L and the feather R; this symbolizes planned speech such as lectures. The base of the feather R curls to the right up around toward the back of the catcher's mitt S; this symbolizes the reading of informal notes which bear some stylistic similarity to informal speech.

BEST COPY AVAILABLE

The lines of various thickness from IA (Interactive) and NP (Non-participatory) suggest the possibility that different kinds of listening might have different relationships to L2 skills.[14] [15]

Bottom line: a fuzzy dichotomy of listening. A number of loosely related concepts have been introduced in this section by language analysts writing from different perspectives. Taken together, these concepts suggest the possibility of making a fuzzy dichotomy between different types of listening. The full elaboration of such a dichotomy is beyond the scope of this paper. However, the ideas presented provide a transition to the work of two NL researchers who have analyzed listening into three component traits.

Bostrom and Waldhart's Three Types of Listening

Bostrom and Waldhart (1981) are the authors of the Kentucky Comprehensive Listening Test mentioned earlier. They identified at least three types of listening behavior, which they call short-term listening, interpretive listening, and lecture or long-term listening.

Short-term listening. In the first part of the test, the examinee hears a series of numbers or letters, sometimes accompanied by background noise. He/she is immediately thereafter prompted to answer a question about the order of the numbers or letters in the series. The examinee must respond *immediately* after the prompt. The authors calls this "short-term listening" (STL).

In the second part of the test, the examinee hears again hears a series of numbers or letters, but no background noise. He/she is prompted to answer a question about the order of the numbers or letters in the series *only after an interval* of 20 to 50 seconds after the last number or letter in the series is presented. The authors call this short-term listening with rehearsal" (STL-R).

Interpretive listening. In the third part, the examinee hears successive parts of a dialogue consisting of very brief interchanges. It is apparent from nonverbal audio and situational clues that the speakers sometimes say one thing and mean something else. The examinee must answer questions about the intent of the speakers by choosing from very brief multiple choice options. The authors call this interpretive listening.[16]

---

[14]This diagram should be interpreted with caution. For example, Figure 1 does not account for certain plausible assumptions about early language learning. One such assumption would be that phonological coding ability, grammatical sensitivity, and ability to acquire vocabulary play a major role in early language learning. These predictors might predict globally across skills. This section has only suggested some nonspecific intuitions about what kinds of predictors might be represented by IA and NP. These ideas have not been specified well enough here to try to identify IA and NP with any of the standard reference factors in the mental testing literature. For further discussion on the concept of different variables being important at different stages of language acquisition, see references by Higgs, T., Clifford, R. (1982), Upshur, J, Homburg, T., (1983), de Jong (1994).

[15] It is interesting to note that multi-method, multi-trait analyses of language skills often find clear trait differences in the case of speaking and reading, but tend to find method and trait confounded in the case of listening. One reason for this kind of confounding might be that an interview method of measuring listening might tap the "S-side" of listening, while a multiple-choice test might tap the "R-side" of listening. Thus Figure 1 might offer some insight into the kind of data found in the reference by Dandonoli, P., Henning G. (1990).

[16]This is a concrete example of a kind of test that might measure a kind of L1 behavior that is closer to "interactive" listening than "noninteractive" listening. However, a much broader sphere of infuence is assigned to interactive listening in Table 1 as a whole, much broader than this one test of "interpretive listening" would measure. Identifying this test with this broad concept of interactive listening would probably go beyond the specific intent of Bostrom and Waldhart.

Long-term listening. In the fourth part of the test, the examinee hears a lecture that is approximately 1500 words in length and must thereafter answer multiple choice questions on the lecture. The examinee is not allowed to take notes. The authors call this lecture listening or long-term listening.[17]

An elaboration of Bostrom and Waldhart introducing the concept of "native authentic listening"

Introducing a concept to elaborate on Bostrom and Waldhart's work. Figure 2 uses visual metaphors to portray relationships between the three listening factors found by Bostrom and Waldhart and another concept I will introduce--"authentic native listening."

Hypothesizing an upper anchor for the ILR listening scale. This new concept itself needs to be elaborated. We need to explain why we as FL researchers have a warrant to use this concept. "Native authentic listening" (NAL) is an extrapolation from a FL learning context to a NL context. We are extrapolating to what a "native listener" would be able to do if he had no need of language instruction. I must consider the construct to be an elaboration on my part because FL researchers like myself devote almost all of our attention to the kind of "authentic listening" that language *learners* with various lesser levels of skill can perform. That is, we don't devote much attention to analyzing, diagnosing, and remediating what native speakers can in some sense already do[18] The term has significance to us not because we observe, think and write a great deal about the concept in our own FL research literature, but because we find it a useful icon for anchoring the end point of the proficiency scale (the theoretical ultimate goal of FL instruction) rather than an object of intensive study in itself.[19]

NAL portrayed as a circle inside a triangle in Figure 2. "Authentic language" (NAL) is represented by a circle inside a triangle. NAL is a set containing "authentic listening" tasks and is located inside the circle. The members of the set are "tasks" and not isolated words and grammatical constructions. This implies that each NAL task consists of a binary relationship involving (1) an authentic NL goal for listening and (2) an accompanying authentic NL text.

Inside the Circle. The members of this set of tasks (defined above as binary relationships) are in different locations inside the circle. These various member tasks are at different distances from the corners of the triangle.

Pure traits portrayed as corners of the triangle. The corners of the triangle represent pure traits (PT) roughly analogous to the traits that Bostrom and Waldhart identified. The metaphor intended here is that various tasks within the circle may require different combinations and weighting of PTs. The combination and weighting of PTs required for individual tasks corresponds to distances from the corners of the triangle.

---

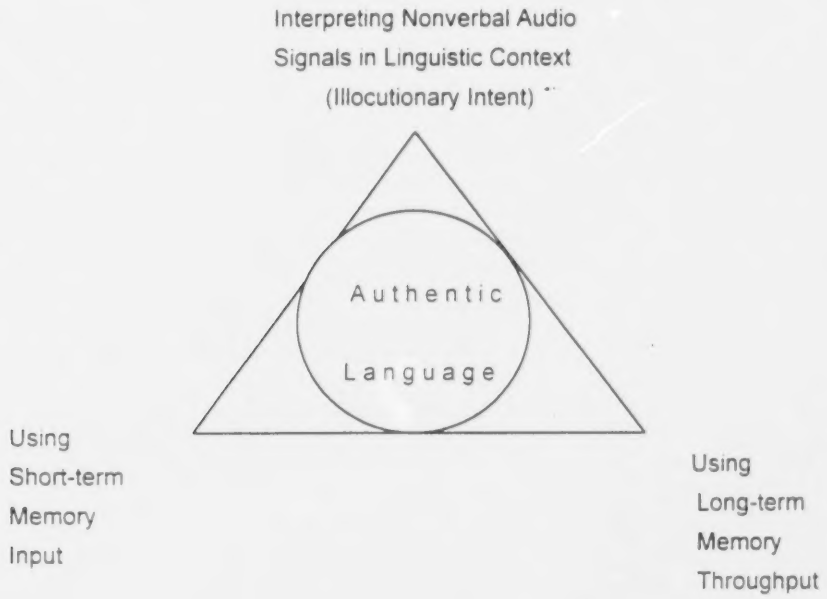
[17]There is also justification for citing this test as an example of "nonparticipatory" listening that is more closely related to reading than speaking.

[18] Consider the following examples of research interests that are seldom found in the FL literature: (1) individual differences in coping ability of native listeners in situations where a speaker introduces new information too quickly for the NL to relate the new information to previously presented ideas, and (2) the kind of notetaking strategies a NL employs in a lecture situation with the intention of later reconstructing and studying the lecture content, in cases where the lecturer presents too many ideas for the NL to follow in real time.

[19]The intention of this extended explanation is to make it easier for those of us in the ILR camp to better communicate with scholars with other research interests by being clearer about our own background interests and thinking.

Why Pure Traits Lie Outside the Circle. The corners of the triangle themselves lie outside the domain of circle that represents NAL. This is a metaphor that has a purpose. It suggests that PTs may predict acquisition of proficiency by language learners without being an NAL task themselves. For example, memory span for letters and numbers is not really a task that belongs to NAL, because NLs seldom make it a listening goal to remember the location of numbers in a string; they don't have any real need to do this as part of their daily life. Nevertheless, memory span for letters *may* predict foreign language proficiency. It is an open question, and an important question, whether a test based on PTs or one based on NAL is a better predictor for the purpose of language aptitude. A broad variety of psychometric and practical issues may bear on the answer to that question.

The scope of these issues is large enough to preclude much discussion of them at this point in this paper. I will return later to the subject of PTs and NAL, and give examples to illustrate the points made above. Before doing that, I want to prepare the ground by addressing yet a third perspective for viewing listening comprehension (in addition to the predictive perspective and the linguistic content perspective). Hopefully, this third perspective will make the examples more cogent.



NATIVE LISTENING

FIGURE 2

Summarizing discussion of predictive perspective and language content perspective.

I conclude this section on the linguistic content approach by expressing another hope. My hope is that the audience perceives there is some connection between one's research background and previous conception of the term "listening" and the number and type of predictors one expects to find under the general rubric "native listening." If that hope is justified, I am ready to present NL from the perspective of cognitive models.

The Last of Three Complementary Approaches in the Literature on Listening Comprehension: the Perspective of Cognitive Models

Introduction

I have two motivations for introducing the topic of cognitive models. One reason is the prominence of the concept in the recent literature. The other reason is more personal. I will start by elaborating my personal interest.

Personal Perspective

I have been struck by the seeming paradox between native listener performance on certain listening tasks involving short simple texts and certain other tasks involving long complex texts. In some cases, the native listener will accomplish the task with the long text much more easily than the task with the short text. I will provide a concrete example later. However, I think the example will be easier to understand if I first make use of an analogy to prime the pump. One element in the analogy is the contrast between the native performance on short and long texts. The other element in the analogy involves computer data bases.

If one has a very large data base with a large number of fields, one can create a targeted set of successive queries that quickly selects three or four cases out of 1,000,000 records that have the exact elements desired. On the other hand, if for some reason it is impossible to use an appropriate query, it can be difficult to find a few records in a much smaller data base.

Historical Perspective

Back to the black box. This example about the role of data base queries suggests a path to move from my personalized perspective to a broader perspective. The broader perspective involves the historical development of cognitive models, including models of listening comprehension. There has been a considerable evolution in the past sixty years from the heyday of radical "black box" behaviorism to current day trends in cognitive psychology. A half century ago, many prestigious mathematical psychologists were loosely associated with the behaviorist school. The radical behaviorist school suggested that if we patiently allowed the mathematicians to analyze data on stimulus strength impinging on the black box, response time, and response strength emanating from the black box, their school would eventually explain complex behavior.[20]

There's somebody in my black box. By the 1960s, many prestigious mathematical psychologists had decided to jump ship. These mathematicians had realized that the data about the responses from the black box don't make much sense unless one takes into consideration not only (1) what the organism in the black box must have known before the stimuli came in; but also (2) what the goals of the organism were when it was learning what it now knew; and even beyond that, (3) still more information about what the inside of the organism in the black box must have looked like all the while.[21] Deprived of the prestige mathematicians had contributed to their stimulus-response theories, the radical "black box" behaviorist school no longer had the ability to attract much attention with their own ideas on complex verbal behavior nor to inhibit other ideas from being developed.

---

[20]The progression of thought in the behaviorist school can be traced in the references by Watson (1924); Hull (1943); Skinner (1957). (The classic and decidely antibehaviorist opposing response to the Skinner reference comes from the field of *linguistics*; see reference by Chomsky, N., [1959].)

[21] A continuous process of evolution is evident from the series of references by Hull, C. (1943); Norman, D., (ed.), (1970), Norman, D. and Rumelhart (1970), Greeno, J. (1970), Montague, W. (1977).

Time to talk about different types of memory. Thus, the first evolutionary step occurred when the mathematicians gave a new breed of psychologists permission to hypothesize on what was inside the black box. At first the hypotheses were relatively simple. There had to be a short-term memory, a long-term memory, and some sort of active working memory where a goal-setting executive transformed information from the outside to fit in with previously learned information from long-term memory.

The computer metaphor. The next evolutionary step occurred when individual researchers began to furnish the black box with any additional construct that helped them explain any of their own behavioral data. This was important because technicians in other fields were making progress in fields such as computer data bases, expert systems, and artificial intelligence. All these developments contributed a new source of metaphors to describe the furniture inside an increasingly transparent "black" box.[22]

Introspection returns to favor. The final evolutionary step occurred after introspective (and retrospective) techniques such as think-alouds returned into favor and became familiar instruments in the cognitive psychologists' tool box. Nowadays investigators commonly use language borrowed from the field of data processing to both describe and elaborate introspective and retrospective data.[23,24]

This last evolutionary step provides the context for me to return to my personal concern with tasks involving short and long listening texts. I will now provide concrete examples to use in a think-aloud. I intend to then use the retrospective data from the think-aloud to construct an analogy with a multimedia database.

Concrete Examples

The first example on the following page is a listening task with a short amount of audio text.

The second example is a listening task with a large amount of audio text. The tiny subscript numbers serve only to identify the sentences in the text for subsequent discussion.

Preliminary discussion of the two passages

The two passages are printed on the following page.

A seeming paradox. Small scale trials indicate that native speakers find the second task easier to perform than the first task. Yet the text for the second task is much longer. It also has a variety of features that might confuse a foreign language learner with little proficiency--such as idioms, reasonably complex grammar, and somewhat culture specific content. If we remember the metaphor of the vast ocean and the vast variance, we might suspect that there are some interesting things to consider about the second passage. These interesting things could correspond to the residue left after our hypothetical ocean evaporated.

---

[22]The progressive development of computer analogies can be traced in the references by Anderson, J., Bowen, G. (1974), Findler, N., (1979), (ed.), Cermak, L. Craik F., (1979), Kolodner, J. (1984). In turn, these computer analogies were intellectually compatible with development of "spreading activation" and related "connectionist" theories in references by Collins, A., Loftus E. (1975); and Cottrell, G., (1994).

[23]The reference by Hintzman (1987) provides a balanced historical overview of the competition and interaction between behaviorist and cognitivist schools of psychology., and summarizes in accessible form the path of evolution represented by the references in footnotes 16-18.

[24]See reference by Faerch, C., Kasper, G. (1987) concerning use of introspective techniques in second language research. Think-aloud techniques also play a role in the language learning strategies literature. See references by Wenden, A., Rubin J., (ed.), (1985),Thain, J., Lett, J., (1991).

AUDIO EXAMPLE 1

Listen to the following series of numbers and be ready to answer a question about the numbers:

024
252
306
408
503

What was the third number presented?

AUDIO EXAMPLE 2

Listen to the following text and be ready to answer a question about the text.

'Let's look at my schedule before I give you a number to call and tell you when you should call about your file. 'If I'm in a meeting or in someone else's office, people will be around nipping at my heels, you see, and not only that I may not have my stuff at hand to talk to you.

'From 9 to noon, everybody including me too, will be at extension 463, that is in theory, but we'll all be behind closed doors at the contract award board.

'From 1 to 3, I'll sneak back to my files at my old office at extension 654, where every one is on leave anyway. 'I to 3 at 654, make a note.

'From 4 to 6, we'll all be back at the Contract Approval Office at extension 625, all of us huddling together to tie up all the loose ends from the morning again, so if the unexpected happens and you can't get me earlier in the afternoon, this is a last resort to call 625 then.

When should you start trying to call me and at what number?

The need to focus. In the first task, the listener probably attempts to hold previous linguistic input from the speaker in his memory in its *original* form, while the speaker continues to provide new input. In this task, the listener (L) *might want* to identify which input is more important and which less important, but the structure of the task gives L *no* opportunity to do so. If L only had a goal that enabled L to decide which input deserved more attention, L might be able to make the important input *more* salient in L's own mind than any less important input that might come later. Unfortunately, L has no clue as how to accomplish this, and thus has no way of preventing later and less important information from driving what ultimately turns out to be important information from L's working memory. If L were to give a list of appropriate verbals and verbal combinations that describe what L would *like to do, but can't do* in this task, that list might include such words as rehearsing, activating/maintaining, focusing, and attending.

Having a goal helps. In the second task, the listener will probably quickly give up on holding most of the input in its original form. Instead L quickly realizes the task is structured in such a way that L can almost immediately define a goal and begin to assimilate important information into larger cognitive structures. The cognitive structures will comprise an interlocking set of interpretive schemata. The seed template for the larger structures existed in some sense in L's long-term memory before the listening task began. Such templates were based on the L's broad past experience.

AUDIO EXAMPLE 2

Listen to the following text and be ready to answer a question about the text.

[1]Let's look at my schedule before I give you a number to call and tell you when you should call about your file. [2]If I'm in a meeting or in someone else's office, people will be around nipping at my heels, you see, and not only that I may not have my stuff at hand to talk to you.

[3]From 9 to noon, everybody including me too, will be at extension 463, that is in theory, but we'll all be behind closed doors at the contract award board.

[4]From 1 to 3, I'll sneak back to my files at my old office at extension 654, where every one is on leave anyway. [5]1 to 3 at 654, make a note.

[6]From 4 to 6, we'll all be back at the Contract Approval Office at extension 625, all of us huddling together to tie up all the loose ends from the morning again, so if the unexpected happens and you can't get me earlier in the afternoon, this is a last resort to call 625 then.

When should you start trying to call me and at what number?

I eventually want to retrace my steps in the previous paragraph, and illustrate why a multimedia data base is a good analogy of the process I am describing. However, I will first prime the pump by briefly elaborating on the function of the larger cognitive structures mentioned in the previous paragraph.

Activating important information and forgetting the rest. The larger cognitive structures will accomplish more than merely assimilating the original information. They will also (1) assimilate succeeding pieces of information that are important in terms of the goal, (2) keep the important information active in working memory, (3) deactivate less important information. Furthermore, the effort required to keep the larger cognitive structure alive in working memory will place less load on the listener's cognitive resources than would a corresponding effort to preserve isolated pieces of information in memory. The new structure will help the listener (1) fill in the gaps beyond what the speaker has explicitly said, *and* (2) "edit out" (into an inactive state) some unimportant things that the speaker actually did say. If L were to give a list of appropriate verbals that describe what L is able to accomplish in this task, that list might include such words as elaborating, interpreting, activating/absorbing, and inferencing.

The Active Listener and the Analogy of a Multimedia Database

Now I can retrace my steps and address the question of why a multimedia data base is a good analogy for what is happening in the second task.

(1) Upon hearing the first sentence in the text "Let's look at my schedule before I give you a number to call and tell you when you should call about your file.", L consults the "data base" under a field named GOALS, and finds a template that matches the input. This template probably tells L to be ready to conduct another search based on fields such as TIME, LOCATION, PHONE EXTENSIONS, FILES, and SPEAKER GOAL to match the expected input.

(2) Upon hearing the second sentence L suspects L should be ready to take any further input and conduct a major sort on PERSONS and a minor sort on LOCATION and PHONE NUMBER, with two intentions in mind. The first intention is to deactivate any piece of incoming information in which more than one PERSON is present. The other intention is to concentrate on any record in which the speaker is the PERSON. In addition, L infers that L should be ready to take the LOCATION and PHONE NUMBER fields of the remaining records and be ready to run major sorts on these fields with minor sorts on SPEAKER INTENT, TIME, and FILES.

AUDIO EXAMPLE 2

Listen to the following text and be ready to answer a question about the text.

Let's look at my schedule before I give you a number to call and tell you when you should call about your file. If I'm in a meeting or in someone else's office, people will be around nipping at my heels, you see, and not only that I may not have my stuff at hand to talk to you.

From 9 to noon, everybody including me too, will be at extension 463, that is in theory, but we'll all be behind closed doors at the contract award board.

From 1 to 3, I'll sneak back to my files at my old office at extension 654, where every one is on leave anyway. 1 to 3 at 654, make a note.

From 4 to 6, we'll all be back at the Contract Approval Office at extension 625, all of us huddling together to tie up all the loose ends from the morning again, so if the unexpected happens and you can't get me earlier in the afternoon, this is a last resort to call 625 then.

When should you start trying to call me and at what number?

(3) Upon hearing the third sentence, L carries out the planned queries, and deactivates the information because the PERSONS field does not match.

(4) Upon hearing the fourth sentence, L carries out the planned queries again, and saves LOCATION, PHONE NUMBER, TIME, and FILES from the input and still has resources left to check the input against SPEAKER INTENT.

(5) Upon hearing the fifth sentence, L verifies SPEAKER INTENT, and activates the follow record: LOCATION (my old office), PHONE NUMBER (654), TIME (1 to 3), FILES (Present), SPEAKER INTENT (Helpful toward meeting listener goal), and GOAL (know where and when to call about file). L will now check any incoming information against this record and deactivate any nonmatching record.

(6) Upon hearing the sixth sentence, L is ready to deactivate incoming information to prevent interference with the previously validated record. This is because the information in the sixth sentence doesn't match all the fields in the previously validated record, (e.g. LOCATION (Contract Award Office), FILES(Inferred to be absent), SPEAKER INTENT (busy solving another problem). By this time L could have forgotten the first phone number because L had already deactivated it. L is also ready to place a priority on rehearsing the record with TIME(1 to 3) and PHONE NUMBER (654), with secondary priority on remembering the last PHONE NUMBER (625), which matches only on SPEAKER INTENT (gives number as last resort).

(7) At this point the test question is given. As soon as L verifies that the activated record is the answer to the question, L fine tunes the GOAL to (provide answer), provides the answer, and deactivates all other information.

(8) There is another field in L's data base that will be activated during this conversation. However, the input matching against this field cannot be localized to a single sentence. If one omits the words and simply hums the discourse intonation, one finds that the intonation itself gives a strong indication where the most important information is.

<u>Lessons to be Learned from these Two Passages</u>

Before proceeding, I will summarize what we can learn from the two passage examples:

Try it, you'll learn something. First of all, I concede to skeptics who think I stacked the deck with these examples to make a rhetorical points that they are right, and I will proceed to make those very rhetorical points. However, I do suggest that interested readers attempt small-scale experiments like the one described above to convince themselves from their own experience that the variables described do play an important role in NL.

Useful database analogy. The data base analogy has been helpful in illustrating that features other than vocabulary, grammar, and passage length can affect NL comprehension. On the other hand, a nonnative speaker with a lesser level of proficiency might have been distracted by some of the very parts of the passage that helped the NL perform the task.

Pedigree of database analogy. I chose to use the analogy of using a database to show how a listener might elect to select certain information and ignore other information. Those familiar with other connectionist approaches might correctly think that my informal analogy has some parallels with these approaches. In brief, a connectionist approach suggests that all the various elements (words, inferred pragmatic goals, grammar, intonation) at *different* levels of linguistic (and perhaps some metalinguistic and nonlinguistic) structure in the spoken input are involved in interpreting an incoming message.[25] They are involved in the sense that they all get to "vote" on what kind of interpretations make sense in terms of the intent of the incoming message. Interpretations that are "voted" as plausible are activated and implausible interpretations are ignored. Activated interpretations provide the context for interpreting the input that follows. Certain elements are more likely to be "connected" or "associated" with each other by context. One can visualize a number of different "images" to represent this kind of "connection:"

(a) In my data base analogy a series of queries scored "hits" or "matches" that influenced successive searches.

(b) Another image might be that "connections" that are stronger support each other (vote for each other) in context and "veto" other less plausible connections.

(c) Another image might be that "connections" that are inherently more plausible in context are awarded more votes and outvote other possibilities.

Forerunners of contemporary connectionist approaches include Collins and Loftus' (1975) theory of spreading activation and Anderson's (1983) adaptive control of thought (ACT).

Recent applications of similar models in artificial intelligence have succeeding in producing machines that can carry on a surprisingly natural conversation within certain limited topic domains. This success seems striking enough to lead me to speculate further on the kind of cognitive abilities required for comprehension skills.

Not just a database, but a multimedia database. I have suggested an analogy be made between the listening process and a *multimedia* database--not just an ordinary database. In order to make this analogy clear, I will elaborate on some of the characteristics of a multimedia database. In a multimedia database, elements might be in text form for some fields, but in the form of video or audio for other fields. The user of such a database might have the capability to inspect the text fields and at the same time call upon peripheral devices to view or listen to the audio and video elements in other fields. This suggests an analogy to the listening process.

---

[25]The metaphor that "a listener actively uses a database to process ongoing discourse" is also compatible with the assumption that the NL tacitly assumes and proactively employs Grice's (1975) maxims to help infer linguistic and discourse structure at every linguistic level, especially the pragmatic level.

The analogy would involve mental processes during listening in which "nontext" elements such as (1) voice affect and (2) intonation patterns could be grouped together under "fields" to be searched. The NL would conduct queries in order to choose matching interpretive schemata to focus his/her ongoing listening process and to deactivate irrelevant schemata during subsequent listening. Just as real mechanical peripheral devices have performance limitations that can be objectively studied, I would hope that connectionist mental models would provide a basis for studying the characteristics and limitations of mental subsystems contributing to listening comprehension. In addition, the mental measurements specialist may find connectionist models suggest hypotheses as to what measures are appropriate to predict and measure comprehension ability and language acquisition. For example, they might speculate that measures testing the processing of vocal elements less directly involved in lexical processing may provide sources of variance distinct from those measures typically associated with strictly lexical processing. This analogy thus suggests the possibility that NL testing should use two distinct listening measures: a "lexical focus" listening measure and a "voice focus" listening measure.[26]

Those that have nothing to seek take longer to find. Real-life database users know well the frustration caused when they try to find a certain single record in a large database file, but don't have a clear idea of what query to use. Sometimes they have to just give up and turn their attention to more pressing business. This familiar experience from the computer world may have a parallel in listening comprehension. Spearitt (1962) administered a large number of listening comprehension measures along with other cognitive tests. He found that tape-recorded tests with such names as Illogical Grouping and Haphazard Speech loaded on a memory span factor.[27]

Spearitt's findings tie in with several other ideas presented in this paper. After our experiment with the short text and the long text, I suggested that the presentation of the shorter text did not allow the listener the opportunity to establish a goal in time to chunk the important input into a larger cognitive structure. It is reasonable to suppose that a longer memory span would give a listener a little more time to hold input in short-term memory before deciding how to chunk it into an appropriate structure.[28]

The argument in the preceding paragraph suggests that we can add a "memory span" variable to the "lexical focus" and "voice focus" variables mentioned above. This is a conclusion similar to the one Bostrom and Waldhart reached through a different route, when they established a distinction between short-term listening, interpretive listening, and long-term (lecture) listening. Of the three traits, only long-term listening seems to have something in common with the measures presently included in the ASVAB and DLAB at this time.

Conclusions and recommendations concerning NL measures

Our review of NL tests and of the literature on listening has enabled us to come to some tentative conclusions. However, since we at DLI don't have much experience in actually writing NL tests. We would like to seek out the opinions and help of experts who have had more practical experience. For this

---

[26] References by Doff, A., Jones C. (1980) and Haycraft, B., Lee, W. (1982) are basic ESL conversational course materials, but with a special twist that may give the reader a hint of some of the kind of skills might be involved in "voice focus" listening measures.

[27] See Carroll's (1993) reanalysis of Spearitt's data set one one of the series of diskettes accompanying Carroll's recent book cited in this reference.

[28] A variety of other studies have addressed a number of relationships between memory span, speed of auditory closure, listening to distorted or illogical speech, and listening to speech with background distractions. The results of these studies seem to be influenced by the variety in testing measures employed and by the specific populations chosen. See references by Karlin, J. (1942), Stankov, L., Horn, J., (1980), Horn, J., Stankov, L. (1982), and related comments by Carroll, J. (1993).

reason, as I present each tentative conclusion. I will also identify areas in which we at DLIFLC might benefit from the expertise of other scholars.

Decision criteria for evaluating alternatives

A good starting point is to list the criteria for evaluating alternative L1 listening test types for inclusion in an expanded language aptitude battery. It is hard to improve on Henning's (1987) largely self-explanatory list of criteria for evaluating language tests, which I quote below:

Purpose of the test:	test validity
Characteristics of examinees:	test difficulty
Precision and accuracy:	test reliability
Suitability of format and features:	test applicability
The developmental sample:	test relevance
Availability of equivalent or equated forms:	test replicability
Scoring and reporting:	test interpretability
Cost of test procurement administration and scoring:	test economy
Procurement of the test:	test availability
Political considerations:	test acceptability

I might add two other criteria relevant to our plans to expand the current DLAB: (1) since tests to be retained from the old DLAB already require 75 minutes to administer, it is undesirable for the total test administration time of an expanded DLAB should exceed two hours; and (2) in order to use the total administration time wisely, DLI would like to avoid adding measures that duplicate any part of the current DLAB or ASVAB, the two screening batteries used to select students.

The above list of criteria gives an idea of DLI concerns. A complete evaluation of NL tests in terms of all these criteria is far beyond the scope of this paper.

Pure traits (PTs) vs. Native Authentic Listening (NAL): A Quick Scan of Current NL Tests as Models

I made a distinction earlier between measures of PTs (pure traits) and NAL (Native Authentic Listening). The concept of measuring PTs derives from a tradition in mental measurements that places a high value on defining minimally intercorrelated traits, --sometimes even at the seeming expense of ecological or face validity in test content. I noted that FL methodologists see NAL as a *theoretical* ideal (in terms of face validity), because they can equate NAL with the upper anchoring point for the ILR proficiency scale for *FL* listeners. The diagram presented earlier in Figure 2 and the accompanying explanatory text explained the relationship between PTs and NAL.

I left the question open as to whether PTs or NAL were the most appropriate measures of NL comprehension as a predictor of L2 proficiency.

Table 2 attempts to list the *number* (expressed by a digit in large type) of instances in which each of the reviewed NL tests contain (1) item types that measure PTs, (2) item types that measure NAL, and (3) item types that are on the borderline between PTs and NAL. The table serves to roughly quantify the occurrence of these types of items in these tests. It is hazardous to draw detailed conclusions from this table, because it does not furnish a very precise categorization of item types. The main conclusion that can

TABLE 2
NL LISTENING TEST ITEMS: PTs or NAL?

Name of Test	Number of distinct item types in each test that tend to measure PTs rather than NAL	Number of distinct item types in each test that measure on the border of NAL, and thus tend somewhat toward measurement of PTs	Number of distinct items types in each test that clearly measure NAL, not PTs
Watson-Barker		**2** !Lecture listening !Emotive listening	**3** !Conversations !Instructions/Directions !Listening for Content
Kentucky Comprehensive Listening Test	**1** !Short/-term memory (2 types)	**2** !Lecture listening !Interpretive listening	
Carleton University Test	**1** !Lecture listening as bootstrap to library research		
NTE Communicative Skills Listening Test		**2** !Interactive situations involving empathic listening !Lecture listening to extended passages on educational topics	**1** !Variety of listening situations especially school situations without strong cognitive or emotional load
NTE School Guidance and Counseling	**1** !Counseling situations involving empathic listening		
Brown-Carlsen	**1** !Immediate recall	**1** !Lecture listening	**1** !Miscellaneous other item types
STEP	**1** !Immediate recall	**1** !Lecture listening	**1** !Miscellaneous other item types

be drawn[29] from Table 2 is that some of the item types measured on NLs are more like measurements of PTs than NAL[30], some of the item types clearly measure NAL, and some are on the border line (the edge of the circle in Figure 1.)[31] Thus, a survey of NL test item types does not in itself give any guidance as to whether one should proceed with a PT measurement approach, an NAL approach, or something in between.

The next two sections deal with the kinds of considerations involved in using PT test content and NAL test content in NL tests used as aptitude tests for predicting FL proficiency.

Measures of PTs as FL aptitude test measures

Three PTs were identified earlier in Figure 1. They involved short-term memory, long-term memory, and interpretation of nonverbal audio signals.

Relation of PT measures to current and future ASVAB. PT measures of long-term memory may tend to share some variance with ASVAB tests that are associated with cognitive and verbal achievement. Furthermore, although there is no short term memory test on the current ASVAB, working memory tests that tap similar abilities have been proposed for inclusion in ASVAB. On the other hand, nothing in current or projected ASVAB versions will test nonverbal audio signals.

Nothing in the current DLAB seems to compare to any of the three PTs.

Using PTs measures of long-term memory and lecture listening measures: choice of content areas. Performance on long-term or lecture listening tasks is facilitated when a listener has access to content-area schemata for the subject areas represented in the listening texts. Depending on the circumstances, knowledge of almost any content area schema acquired prior to L2 study could potentially be useful in L2 listening, especially after the L2 listener has surmounted initial phonological, grammatical, and lexical hurdles.

However, it is likely that some broadly conceived content-area schemata would be particularly relevant:(1) international and cross-cultural communication; (2) issues of sensitivity to international and cross-cultural differences; (3) international business, political, cultural, and military cooperation (or rivalry), (4) cross-cultural technological transfer (or maintenance of technological secrecy); and (5) comparative political science. On the other hand, one could easily name a number of content area schemata that

---

[29] My sources of information were test information brochures, published information, and personal communications, rather than a detailed review of the physical contents of each test. In some cases, I have combined what the publisher considered two or more item types into a single item type to more simply fit into my classification scheme.

[30] (The reader may wish to simultaneously refer to Table 2 below and to Figure 2 [which was presented earlier] to follow this footnote.) I identified short-term memory tasks and immediate recall tasks as PT measurements. They fall outside NAL near the "short-term memory" corner of the triangle. I identified the Carleton University task with PT measurement, since general academic ability is important in carrying out that task. This test falls outside NAL near the "long-term memory throughput" corner of the triangle. Similarly, the kind of listening in the "School Guidance and Counseling Examination" is located near the "illocutionary interpretation of non-verbal signals" corner of the triangle.

[31] In cases, where the trait specialization is not as striking as in the previous footnote, I locate lecture listening on the border of NAL and tending toward "long term-memory," whereas I place "emotive listening" and "interpretive listening" on the border of NAL and tending toward the "interpretation of nonverbal illocutionary intent" corner of the triangle.

(especially if narrowly interpreted) would probably be less useful in learning to listen in a language--schemata for abstract mathematical concepts, American sports history, local American building codes, and personal histories of American radio and television entertainers, come to mind as examples.

This suggests that there is a tradeoff to consider in the choice of topic areas for lecture listening tasks. The broad nature of potential applications of foreign languages implies that the choice of topics should be relatively general, but the very nature of career interests of FL listeners suggests that some topics are more appropriate than others. This issue is not only salient for the task for designing a listening component for a FL aptitude test. The publishers of NL tests generally have an "*occupational content target area*" to guide their selection of content; they probably also have to think about striking a balance between general and specialized topic areas. This is one area in which DLI could probably learn from an exchange of experiences with the writers of the NTE Basic Communication Skills Listening Test, the Watson-Barker Listening Comprehension Test, the Kentucky Comprehensive Listening Test, and the test used by Carleton University.

Using PTs as measure of listening skills involving perception of affect. Just as *content* schemata help the listener understand lectures, it is likely that *situational* schemata help the listener understand audio messages with strong affective overtones. It makes sense to talk about a *situational target area(s)* for a NL listening test including an measure of sensitivity to affect. As in the case of occupational content target areas, the situational target area(s) for a NL listening test used for aptitude prediction could differ from target area(s) of such current NL tests as the NTE Basic Communication Skills Test, the NTE School Guidance and Counseling Examination, the Kentucky Comprehensive Listening Test, or the Watson-Barker Comprehension Test. All of the above tests vary in the number of individual test items, the number of situations, breadth of coverage across situations, item length, degree of context provided, and the extent to which cognitive information and affective information are both presented in the same text.

One concern is that the danger of subjectivity or low reliability in tests that measure mainly sensitivity to illocutionary intent or affect, rather than objective cognitive or semantic information.[32] On the other hand, if tests that measure only affect could be made reliable, these tests could turn out to be a potential new source of variance and predictive power. This is because these tests may not share much covariance with verbal and mathematical factors on ASVAB, or with phonological coding and grammatical sensitivity factors on the current DLAB.

The authors of the NL tests mentioned above had to consider a balance between (1) general and specialized situations; (2) long and short items; (3) items involving cognitive knowledge *and* situational sensitivity as opposed to items in which only situational sensitivity seems to matter; (4) and between alternatives in overall content coverage in test planning. The content coverage in some sense has in each case to be appropriate to the career interests of the potential test examinees and the purposes of the test. Again, this is area in which DLI could probably learn by exchanging experiences with NL testers as to how to select test content appropriate to the career focus of the FL linguist.

Measures of NAL as FL aptitude test measures

It is possible to base discussion of test content solely on NAL, rather than PTs. However, even if all the test content was genuine NAL, one could still suppose that each component item would represent a task that requires some cognitive contribution from each of the PTs. Some item tasks would require greater cognitive contributions from some PTs than other PTs.

For example, NL for certain kinds of instructions and directions could place more of a load on short-term memory than long-term memory or illocutionary sensitivity. Certain other NAL items could easily involve NL tasks that place higher demands on either: (1) affective and situational sensitivity, or (2) cognitive or academic sensitivity.

---

[32]See Bostrom, R. (1990a), p. 19.

Some of the same issues in content selection mentioned above in the discussion of PTs would also thus apply even for a test focused on NAL. From this point of view, DLI could benefit from exchanging experiences with the writers of tests like the NTE Core Battery Listening Test or the Watson-Barker Listening Test. These tests have placed somewhat less emphasis on breaking NL into separate or specialized traits than have some of the other tests listed above.

Bottom Line on NL tests as predictors

Several kinds of L1 listening tests are likely candidates for an FL aptitude battery.

Since DLI doesn't have much experience in actually writing NL tests, our agency could benefit from interaction with NL researchers with interests outside the FL testing field. There has not been a great deal of communication of between the disciplines of FLL and NL research. This is an area where DLI could foster a basic exchange of information concerning research interests and backgrounds between researchers in these two disciplines. Subsequent interdisciplinary exploratory efforts could play a very important role in the revision of the DLAB.

Subject to feedback resulting from such interdisciplinary interactions, I can draw certain tentative conclusions.

General conclusions. It would be best if the addition of a L1 test should not greatly increase the length of the DLAB. A revised DLAB (including both old retained tests and new added tests) should not exceed two hours in administration time. Optimal administration time would be somewhat less than two hours

There should be no copyright or licensing problems that would prevent unrestricted duplication and subsequent administration of tests by the Department of Defense (DoD). DoD would want to retain unfettered controls over the administration and test security of any test added to the DLAB.

As explained earlier in the section on the predictive perspective, DLI should consider adding tests that are different from any test currently used in ASVAB and DLAB, the currently used screening instruments. The rationale for having a different kind of test, is that a different test is more likely to measure something new and not duplicate variance already measured by another test.

Test content. From this point of view, DLI should consider tests of short-term memory and tests focusing on vocal quality or sensitivity to illocutionary intent. These tests might be less likely to duplicate the verbal factor variance found in ASVAB. Tests of such abilities might be designed in such a way to also measure auditory perceptual closure and resistance to distraction and auditory distortion. Alternatively, one could consider separate tests for perceptual closure and resistance to auditory distraction or distortion. Although it is desirable that new abilities be measured, DLI needs to also be concerned with the reliability of potential new measures. Of course, it is doubtful that an unreliable test can contribute much additional predictive power to a revised battery.

DLI should not completely exclude the possibility of adding listening tests that are likely to load on a verbal factor. If we elect to design such tests for inclusion in DLAB, we should consider focusing on occupational and situational content target areas. However, one should also consider including a broad range of content areas corresponding to great number of potential applications of foreign language proficiency.

Exploring Tests of Grammatical Sensitivity in English

A review of the tests of grammatical sensitivity is presented. It is not accompanied by an extensive review of the literature on testing grammatical skills comparable to the review of the NL literature given earlier.

The review comprises both: (1) tests of sensitivity to English grammar, and (2) tests that measure sensitivity to foreign (or artificial) language rules.

In contrast to NL tests (which have never before appeared as parts of an language aptitude battery), some tests of grammatical sensitivity have previously been incorporated as parts of aptitude test batteries.

Tests of sensitivity to English grammar

English Grammar Recognition Test (EGRT)

The EGRT was developed at DLI in 1975. It measures explicit knowledge of grammatical terminology. An example of the type of item found in the EGRT is given below:

> A word that modifies a verb or adjective by expressing time, place, manner or degree is called:
>
> a. intensifier
> b. gerund
> c. adjective
> d. adverb

The Flanagan Expression Test (FET)

The Flanagan Expression Test, published by Science Associates, does not require knowledge of grammatical terminology. It has two parts.

In Part One the examinee must identify whether each of a series of English sentences is correct in terms of grammar or usage. An example of a Part I item is given below:

R W I done the work at home.

In Part Two, the examinee must identify which one of three sentences is the "best" way to express an idea.

> ___Most of Greenland consists of glaciers and barren highlands, and no more than two per cent of the island is inhabited and so it is very sparsely populated.
>
> ___Greenland is very sparsely populated. Barely two-percent of the island is inhabited, the rest consisting of glaciers and barren highlands.

The test as a whole has 50 items and takes a little over five minutes to administer. Thus the test is heavily speeded.

DLI efforts to conduct statistical analysis on the FET have been hindered by the fact that student responses to the FET must be recorded on a proprietary non-machine scorable answer sheet.

Preliminary analyses suggest that a large part of the test variance in our test population might be accounted for by a small number of items measuring case and number agreement.[33]

MLAT Part IV (Words in Sentences)

The Modern Language Aptitude Test (MLAT) is published by the Psychological Corporation. It has five parts. Part IV is designed to measure ability to understand the function of words and phrases in sentence structure, without calling upon knowledge of grammatical terminology. Each item consists of a key sentence with a word or phrase printed in capital letters, followed by one or more sentences with words or phrases underlined and numbered. The examinee is directed to pick the word or phrase in the second sentence or sentence group which does the same thing in that sentence as the capitalized word does in the key sentence. An example of the type of item found in MLAT Part IV is given below.

He spoke VERY well of you.

Suddenly the music became quite loud.
1 2 3 4

Tests that measure sensitivity to foreign (or artificial) language rules.

Pimsleur Part IV (Language Analysis)

The test booklet presents a number of words and sentences in Karbardian (a language spoken in the former Soviet Union), and their English equivalents. From these examples, the examinee must figure out how to say 15 new sentences in Karbardian. The items require the application of the examinee's sensitivity to grammatical systems. The examinee is given twelve minutes to answer 15 items.

DLAB Part III (Foreign Language Grammar)

The examinee's task is to learn some grammar rules of an artificial language and then apply these rules in the translation of short phrases and sentences. The words and sentences of the artificial language are similar in some respects to those of English in pronunciation and meaning but have been transformed by the application of rules of the artificial language morphology and grammar. For each item in the test, (1) the examinee *reads* an English phrase or sentence in the booklet, (2) *listens* to the four *alternative translations* in an artificial language spoken on the test *audiotape*, (3) and marks the correct translation on the answer sheet.

The test is so designed that the examinee is effectively discouraged from using a consistent strategy of "reasoning out" the rules to produce a correct answer. For example, (1) the English sentences to be translated are on a separate page from the rules; (2) the examinee is mentally focused on listening to the audio multiple-choice options on the tape; and (3) the examinee cannot review all the options at the same time because the options are presented in serial order on the test audiotape.

Thus as the test progresses and increasingly more grammar rules are introduced, the examinee must become progressively more dependent on automatic processing of previously presented grammar rules.

---

[33] In all of these items, the noun phrases that govern the agreement include either coordinate or complex noun phrases. Strong individual response differences are found in Part I items with stimulus sentences of the type "The videotape playback shows that each of the men and women notice the thief breaking in the office." It is unclear whether individual student differences in answering these items arise from failure to understand a grammatical rule or its scope of application, or from difficulty in applying the rule due to a combination of test speededness and grammatical complexity of the governing construction.

DLAB Part IV (Foreign Language Concept Formation)

The examinee sees four pictures at the *top* of every page of the test booklet. Each picture is accompanied by a description in *an artificial language* of the object or activity depicted in the picture. Taken together these associated pictures with artificial language text constitute a "linguistic corpus" at the top of the page that an examinee must utilize to find correct answers to test items printed on the *bottom* half of that page.

Each item in the bottom half of the page consists of (1) a *picture* and (2) *four written multiple choice options* in the artificial language.

The examinee must find appropriate analogies based on the information in the corpus *and* the individual item to determine which *option* should be matched with the numbered *picture* for that item. The test is moderately speeded.

A completely different set of pictures in a completely different artificial language is introduced on *each* succeeding page. In order to complete the analogies on each page, the examinee has to determine what type of information is relevant to solve the problems on that page. The needed information might be the main concepts underlying each set of pictures, or the graphemic, morphological, or syntactic similarities between the corpus and the individual options for each item. Thus the examinee must have a sensitivity for what kinds of grammatical, morphological, and semantic analogies are possible in a foreign language to solve the problems represented by each item.

Bottom Line on Grammar Tests as Predictors

General

I have completed a review of some tests of grammatical sensitivity, but I have not yet gone ahead to review the literature and issues related to the use of such tests as language tests.

The review of NL tests and literature might provide a useful model for a follow-up review of grammar tests. As in the case of NL tests, DLI could address the utility of grammar as aptitude tests from three perspectives. I will sketch a tentative idea of the components of such a three-part review below.

Predictive Approach

It would be important for DLI to consider the predictive perspective for grammar tests in much the same I did for NL tests. The goal would be to identify the kind of grammar tests that would be most likely to add another source of predictable variance, and less likely to duplicate variance already measured in the current DoD linguist screening process.

"A Grammar Learning Factor Approach"

The next approach in the review of the NL literature was the linguistic content approach. However, grammatical sensitivity is not itself one of the four language skills, but a factor that cuts across all of the four skills. Furthermore, there has been considerable evolution in thinking and ongoing debate for many years as to the proper role of grammar in language learning, and especially to the contribution and relevance of grammar to language learning at various points of the ILR scale. In a review of *grammar* tests, the second approach might be better named the "*grammar learning factor approach.*"

A section devoted to this approach might identify different skills measured in tests such as the EGRT (knowledge of grammatical terminology and ability to apply such terminology in formally analyzing sentences), MLAT (ability to detect parallel grammatical functions and structures in pairs of sentences), and the Flanagan Expression Test (ability to identify grammatical and stylistic correctness under speeded

conditions). It would be profitable to investigate how foreign language methodologists rate each of these tests in terms of "face validity." Such ratings would no doubt influenced by their own backgrounds in teaching foreign languages and analyzing foreign language acquisition. Such backgrounds, however relevant to foreign language instructional experience, might need to be supplemented by information from a third perspective.

A Cognitive Models Approach

The last approach in the review of NL literature was the cognitive models approach. A parallel approach devoted to grammar tests might focus on experimental psycholinguistic research and studies of computational parsers. Psycholinguistic research of this type might be concerned with human parsing preferences where multiple grammatical clues are present. This type of approach might lead in different directions from the second approach. The second approach, as suggested above, is grounded in classroom language teaching experience rather than formal analysis of the operation of grammatical systems.

Where We Go from Here

Although I have not conducted an exhaustive literature review, I am certain there is an abundance of literature corresponding to each of the three approaches, but no concise synthesis of how the three approaches might relate to the use of grammar tests as language aptitude measures.

I think an intermediate step is needed before DLI develops such a synthesis on its own. DLI should continue to foster an exchange of ideas about the role of grammar in language acquisition and about the role of grammar tests in language aptitude testing. Scholars in the fields of foreign language methodology, psycholinguistics, and cognitive psychology could make valuable contributions to this exchange. Hopefully, these contributions would be a stimulus for DLI to conduct a thoughtful review of the literature at a later time. The intent of this review of would be to evaluate specific types of grammar tests for inclusion in a revised DLAB.

References

Anderson, J. (1983). *The architecture of cognition*. Cambridge: Harvard University Press.

Anderson, J., Bowen, G. (1974). *Human associative memory*. Washington: Hemisphere Publishers.

Arnold, C. (1982). On listening: what does rhetoric have to say to cognitive psychology? In Weimer, W., Palermo D., *Cognition and the Symbolic Process, Vol. 2*, 131-157. Hillsdale, NJ: Lawrence Erlbaum.

Bostrom, R. (1990a). Measuring individual differences in listening. In Bostrom, R., (ed.), (1990) *Listening behavior: measurement and application*. New York: Guilford Press.

Bostrom, R., (ed.), (1990b). *Listening behavior: measurement and application*. New York: Guilford Press.

Bostrom, R., Waldhart, E. (1978). *The Kentucky comprehensive listening skills test*. Lexington, KY: Kentucky Listening Research Center.

Brown, J., Palmer, A. (1988). *The listening approach: methods and materials for applying Krashen's input hypothesis*. London: Longman.

Burgoon, J. (1985). Nonverbal Signals. In Knapp, M., Miller, G., (eds.), *Handbook of interpersonal communication*, p. 344-390. Beverly Hills: Sage Publications.

Byrnes, H. (1987). Features of pragmatic and sociolinguistic competence in the oral proficiency interview. In Valdman, A., (ed.), *Proceedings of the symposium on the evaluation of foreign language proficiency*, p. 167-177. Bloomington: Indiana University.

Byrnes, H. (1984). The role of listening comprehension: a theoretical base. *Foreign Language Annals*, (17), 317-329.

Canale, M. (1983). On some dimensions of language proficiency. In Oller, J., *Issues in Language Testing Research*, p. 188-202.. Rowley, MA: Newbury House.

Canale, M., Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, (1), 1-47.

Carroll, J. (1958). A factor analysis of two foreign language aptitude batteries. *Journal of General Psychology*, (59), 3-19.

Carroll, J., (1993). *Human cognitive abilities:a survey of factor analytic studies*. Cambridge: Cambridge University Press.

Carroll, J. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals*, (1), 131-151.

Carroll, J. (1962). Prediction of success in intensive foreign language training. In Glaser, R. (ed.) *Training Research and Education*. Pittsburgh: University of Pittsburgh Press.

Carroll, J., Sapon S. (1959). *Modern Language Aptitude Test*. New York: Psychological Corporation.

Cermak, L. (1972). *Human memory: research and theory*. New York: Ronald Press Company.

Cermak, L., Craik F., (1979). *Levels of processing in human memory*. New York: Lawrence Erlbaum.

Child, J. (1987). Language proficiency levels and the typology of texts. In Byrnes, H., Canale, M., (eds.), *Defining and developing proficiency: guidelines, implementations, and concepts*, p 97-106. Lincolnwood , IL: National Textbook Company.

Chomsky, N. (1959). Review of *Verbal Behavior* by B. F. Skinner. *Language*, 35.26-58.

Child, J., Clifford, R., Lowe, P. (1993). Proficiency and performance in language testing. *Applied Language Learning*, 4:1, 19-54.

Clark, J. (1986). *A study of the comparability of speaking proficiency interview ratings across three government language training agencies*. Washington: Center for Applied Linguistics.

Clark, J., Clifford, R. (1987). The FSI/ILR/ACTFL Proficiency scales and testing techniques: development, current status, and needed research. In Valdman, A., (ed.), *Proceedings of the symposium on the evaluation of foreign language proficiency*, p. 1-18. Bloomington: Indiana University.

Cody, M., McLaughlin, M. (1985). Situation as a construct in interpersonal communication research. In Knapp, M., Miller, G., (eds.), *Handbook of interpersonal communication*, p. 263-312. Beverly Hills: Sage Publications.

Collins, A., Loftus E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, (82), 407-428.

Cottrell, G. (1994). Connectionist approaches to natural language processing. Article in Asher, R., Simpson, J., (eds.), *Encyclopedia of Languages and Linguistics*, Vol. 2, p 698-706. Oxford: Pergamon.

Crookall, D., Saunders D. (1989). *Communication and simulation: from two fields to one theme*. Philadelphia: Multilingual Matters.

Cummins, J. (1983). Language proficiency and academic achievement. In Oller, J., *Issues in Language Testing Research*, p. 188-202. Rowley, MA: Newbury House.

Dandonoli, P., Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure, *Foreign Language Annals*, (23), 11-22.

de Jong, J., (1994). (untitled guest speaker presentation at Language Aptitude Invitational Symposium.)

Department of Defense (1976). *The English Grammar Recognition Test*. Presidio of Monterey: Defense Language Institute.

Department of Defense (1977). *Defense Language Aptitude Battery*. Presidio of Monterey: Defense Language Institute.

Department of Defense (1985). *Technical supplement to the counselor's manual for ASVAB-14*. North Chicago IL: Military Entrance Processing Command.

Doff, A., Jones, C. (1980). *Feelings: A course in conversational English*. Cambridge: Cambridge University Press.

Dunkel, P. (1991). Listening in the native and second/foreign language: toward an integration of theory and practice. *TESOL Quarterly*, (25)3, 431-457.

Dunkel, P., Henning, G., Chaudron, C. (1993). The assessment of a listening comprehension construct. *Modern Language Journal*, 77(2), 180-191.

Dunkel, P., Pialosi, F. (1982). *Advanced listening comprehension: developing aural and note-taking skills*. Cambridge: Newbury House.

Educational Testing Service (1992). *ETS Test Information Brochure No. 42*, Princeton: NTE Programs: School Guidance and Counseling.

Educational Testing Service (1992). (information pamphlet), *Core Battery Tests*, Princeton: NTE Programs.

Faerch, C., Kasper, G. (1987). *Introspection in second language research*. Cleveden, England: Multilingual Matters.

Fahlman, S. (1981). Representing implicit knowledge. In Anderson, J., Hinton, G., *Parallel models of associative memory*, p. 145-160. Hillsdale: Lawrence Erlbaum.

Feyton, C. (1991). The power of listening ability: an overlooked dimension in language acquisition. *Modern Language Journal*, (75), 173-80.

Findler, N.,(ed.), (1979). *Associative networks: representation and use of knowledge by computers*. New York: Academic Press.

Flanagan, J. (1960). *Flanagan Industrial Tests: Expression*. Chicago: Science Research Associates.

Gardner, R., Lalonde, R., Moorcraft, R., Evers, F. (1985). Second language learning: correlational and experimental considerations. *Language Learning*, 35(2), 207-272.

Gardner, R., Lalonde, R., Pierson, R. (1983). The sociocultural model of second language acquisition: an investigation using LISREL causal modeling. *Journal of Language and Social Psychology*, 2 (1), 1-15.

Glen, E. (1989). A content analysis of fifty definitions of listening. *Journal of the International Listening Association*, (3), 21-31.

Goffman, E. (1983). *Forms of talk*. Philadelphia: University of Pennsylvania Press.

Greeno, J. (1970). How associations are memorized. In Norman, D. , *Models of human memory*, p 257-304. New York: Academic Press.

Grice, H. (1975). Logic and conversation. In Cole, P., Morgan, J. (eds.), *Syntax and semantics: speech acts, Vol. 3*, p. 41-58. New York: Academic Press.

Galvin, K. (1985). *Listening by doing--developing effective listening skills*. New York: National Textbook Co.

Hadley, A. (1993). Proficiency-oriented learning: origins, perspectives, and prospects. In Phillips, J.,Diaz, J., *Reflecting on proficiency from a classroom perspective*. Lincolnwood IL: National Textbook Company.

Hatch, E., Yoshitomi, A. (1993). Cognitive processes in language learning. In Hadley, A., *Research in Language Learning Principles, Problems, Prospects*. Lincolnwood IL: National Textbook Company.

Hatch, E., Shirai, Y., Fantuzzi, C. (1990). The need for an integrated theory: connecting modules. *TESOL Quarterly*, 24 (4), 697-716.

Haycraft, B., Lee, W. (1982). *It depends on how you say it: dialogues in everyday social English*. Oxford: Pergamon.

Heileman, L., Kaplan, I. (1985). Proficiency in practice: the foreign language curriculum. In James, C., (ed.), *Foreign Language Proficiency in the Classroom and Beyond*, p. 55-78. Lincolnwood, IL: National Textbook Company.

Henning, G. (1987). *A guide to language testing*. Cambridge MA: Newbury House.

Higgs T., Clifford, R. (1982). The push toward communication. In Higgs, T., (Ed.), *Curriculum competence and the foreign language teacher*, p 57-80. Skokie IL: National Textbook Company.

Hintzman, D. (1978). *The psychology of learning and memory*. San Francisco: W.H. Freeman and Co.

Horn, J., Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence*, (6), 165-185.

Hull, C. (1943). *Principles of Behavior*. New York: Appleton.

Hymes, D. (1972). On communicative competence. In Pride, J., Holmes, J. (ed.), *Sociolinguistics*, p. 269-293. London: Penguin Books.

Jackson, G., (ed.), (1994). *Final report series for the Language Skill Change Project (Staffing Edition)*. Presidio of Monterey: Defense Language Institute Foreign Language Center.

Jacobs, S. (1985). Language. In Knapp, M., Miller, G., (eds.), *Handbook of interpersonal communication*, p. 313-343. Beverly Hills: Sage Publications.

James, C., (ed.), (1985). *Foreign language proficiency in the classroom and beyond*. Lincolnwood, IL: National Textbook Company.

Janssen, C., Hansen, C., Buck, G., DesBrisay, M., Fox, J., Shohamy, E. (1993). *Writing your own language tests*. (Workshop presented at 1993 TESOL convention.)

Karlin, J. (1942). A factorial study of auditory function. *Psychometrika*, (7), 251-278.

Kolodner, J. (1984). *Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model*. Hillsdale: Lawrence Erlbaum.

Knapp, M., Miller, G., (eds.), (1985). *Handbook of interpersonal communication*. Beverly Hills: Sage Publications.

Kramsch, C. (1987). Response to Heidi Byrnes. In Valdman, A., (ed.), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*, p. 183-187. Bloomington: Indiana University.

Kass, R., Mitchell K., Grafton F., Wing, H. (1983). Factorial validity of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10. *Educational and Psychological Measurement*, (43), 1077-1087.

Krashen, S. (1987). *Principles and practice in second language acquisition*. London: Prentice Hall.

Lett, J., Thain J. (1994). *The Defense Language Aptitude Battery: what is it and how well does it work?* Paper delivered at Language Aptitude Invitational Symposium.

Long, D. (1989). Second language listening comprehension: a schema-theoretic perspective. *Modern Language Journal*, (73), 32-40.

Lowe, P. (1982). *ILR Handbook on Language Proficiency Testing*. Presidio of Monterey: Defense Language Institute.

Lowe, P. (1985). The ILR scale as a synthesizing research principle: the view from the mountain. In James, C., (ed.), *Foreign Language Proficiency in the Classroom and Beyond*, p 9-53. Lincolnwood, IL: National Textbook Company.

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal*, (75), 196-204.

Lund, R. (1990). A taxonomy for teaching second language listening. *Foreign Language Annals*, (23), 105-15.

Magnan, S. (1987). Legal caveats on communicative proficiency testing for graduation requirements or teacher certification. In Valdman, A., (ed.) (1987), *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*., p. 45-48. Bloomington: Indiana University.

MacWhinney, B. (1994). *Psycholinguistic issues in the assessment of the subcomponents of language abilities*. Paper delivered at Language Aptitude Invitational Symposium.

MacWhinney, B. Bates E., (eds.), (1989). *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press.

McLaughlin, M. (1984). *Conversation: how talk is organized*. Beverly Hills: Sage Publications.

Montague, W. (1977). Elaborative strategies for verbal learning and memory. In Bower, G. *Human memory: basic processes*, p 359-436. San Francisco: Academic Press.

Myers, J., McCauley, M. (1985). *Manual: a guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto: Consulting Psychologists Press.

Norman, D. (1970). *Models of human memory*. New York: Academic Press.

Norman, D., Rumelhart, D., (1970). A system for perception and memory. In Norman, D. (1970), *Models of Human Memory*, p. 19-64. New York: Academic Press.

Omaggio, A. (1986). *Teaching language in context*. Boston: Heinle and Heinle.

O' Malley, J., Chamot, A. (1989). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.

O'Mara, F. (1994). Training approaches for reducing student attrition from foreign language training, in Jackson, G., *Final report series for the Language Skill Change Project (Staffing Edition)*, Presidio of Monterey: Defense Language Institute Foreign Language Center.

O'Mara, F., Thain, J. (1994). *Improving the measurement of language aptitude: a psychometric analysis of the Defense Language Aptitude Battery*, paper delivered at Language Aptitude Invitational Symposium.

Oxford, R. (1990). *Language learning strategies: what every teacher should know*. New York: Newbury House.

Petersen, C., Al-Haik, A. (1976). The development of the Defense Language Aptitude Battery (DLAB), *Educational and Psychological Measurement*, (36), 369-380.

Pimsleur, P. (1966), *Pimsleur Language Aptitude Battery*. Psychological Corporation: San Antonio.

Richards, J. (1983). Listening comprehension: approach, design, and procedure. *TESOL Quarterly*, (17), 219-240.

Rost, M. (1990). *Listening in Language Learning*. New York: Longman.

Rutherford, W., Smith M. (eds.), (1988). *Grammar and second language teaching: a book of readings*. Boston: Heinle and Heinle.

Silva, J., White L. (1993). Relationship of cognitive aptitudes to success in foreign language training, *Military Psychology*, 5:2, 79-93.

Skinner, B. (1957). *Verbal Behavior*. New York: Appleton.

Spearitt, D. (1962). *Listening comprehension--a factorial analysis*. (ACER Research Series No. 76). Melbourne: Australian Council for Educational Research.

Stankov, L., Horn, J. (1980). Human abilities revealed through auditory tests. *Journal of Educational Psychology*, (72), 21-44.

Swaffar, J., Bacon S. (1993). Reading and listening comprehension: perspectives on research and implications for practice. In Hadley, A., *Research in Language Learning: Principles, Problems, Prospects*. Lincolnwood IL: National Textbook Company.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives, *Language*, (58), 1-21.

Tannen, D. (1982). The oral-literate continuum in discourse. In Tannen, D. (ed.), *Spoken and written language: exploring orality and literacy*, p 1-16. Norwood, NJ: Ablex.

Thain J., Lett, J. (1991). *The Learning Strategies Project status report and CY92 Plan* (ESR Report 91-02). Presidio of Monterey: Defense Language Institute Foreign Language Center.

Upshur, J, Homburg, T. (1983). Some relations among language tests at successive ability levels. In Oller, J., *Issues in Language Testing Research*, p. 188-202,. Rowley, MA: Newbury House.

Ur, P. (1984). *Teaching listening comprehension*. Cambridge: Cambridge University Press.

Valdman, A., (ed.) (1987). *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*. Bloomington: Indiana University.

Warren W., Nicholas, D. Trabasso, T. (1979). Event chains and inferences in understanding narratives. In Freedle, R., (Ed.), *New directions in discourse comprehension*. Norwood, NJ: Ablex.

Watson, J. (1924). *Behaviorism*. Chicago: University of Chicago Press.

Watson, K., Barker L. (1987). *Watson-Barker Listening Test Video Version*. Auburn, AL: Spectrum.

Wenden, A., Rubin J. (eds.) (1985). *Learner strategies in language learning*. Englewood Cliffs: Prentice-Hall.

Wolvin A., Coakley, C. (1992). *Listening, (4th edition)*. Dubuque IA: Wm. C. Brown.